IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Guillermo Schneider
1/20/22

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with Data Visualization
  - EDA with SQL
  - Build an Interactive Map with Folium
  - Build a Dashboard with Plotly Dash
  - Predictive Analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Dashboard
  - Predictive Analysis Results

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
  - Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problems we want to find answers to
  - What variables most influence launch outcome?
  - Predict launch outcomes

Section 1

Methodology

Methodology

- Data collection methodology:
  - SpaceX REST API
  - Web Scraping
- Perform data wrangling
  - Filtered data to only include relevant Falcon 9 launches
  - Dealt with missing values
  - Treated categorical variables with one-hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

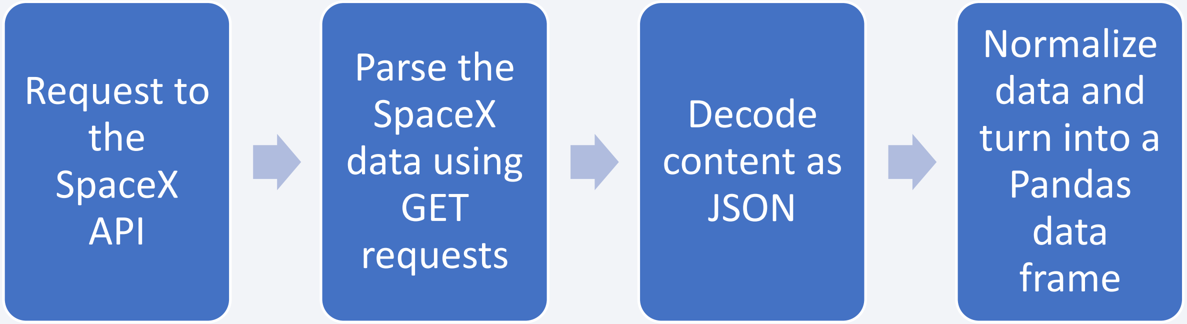
Data Collection

- SpaceX launch data that is gathered from an API, specifically the SpaceX REST API.
- This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- We want to transform this raw data into a clean dataset which provides meaningful data on the situation we are trying to address.

Data Collection – SpaceX API

Variables of interest

- Booster name (and flight number)
- Mass of the payload
- Orbit destination
- Name (and location) of launch sites
- Outcome (and type) of the landing
- Additional rocket factors (Gridfins, cores, landing pad, legs, etc)



Reference: https://github.com/SchneiderG/Python-for-Data-Science-Labs/blob/main/Capstone/1_DataCollectionAPI.ipynb

Data Collection - Scraping

Variables of interest

- Flight number
- Launch site
- Payload mass
- Orbit destination
- Customer
- Launch Outcome
- Date and Time

Request the Falcon9 Launch wiki page from URL → Extract data using Beautiful Soup → Parse HTML tables for variables of interest → Export to CSV

Reference: https://github.com/SchneiderG/Python-for-Data-Science-Labs/blob/main/Capstone/2_Webscraping.ipynb

Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully.
- We grouped landing outcomes to be converted to Classes with “0” as the booster did not land and “1” the booster did land.

Successful (1)	Unsuccessful (0)
True Ocean *(good landing in ocean)*	**False Ocean** *(bad landing in ocean)*
True RTLS *(good landing on ground pad)*	**False RTLS** *(bad landing on ground pad)*
True ASDS *(good landing on drone ship)*	**False ASDS** *(bad landing on drone ship)*
	None ASDS *(failure to land)*
	None None *(failure to land)*

Reference: https://github.com/SchneiderG/Python-for-Data-Science-Labs/blob/main/Capstone/3_DataWrangling.ipynb

EDA with SQL

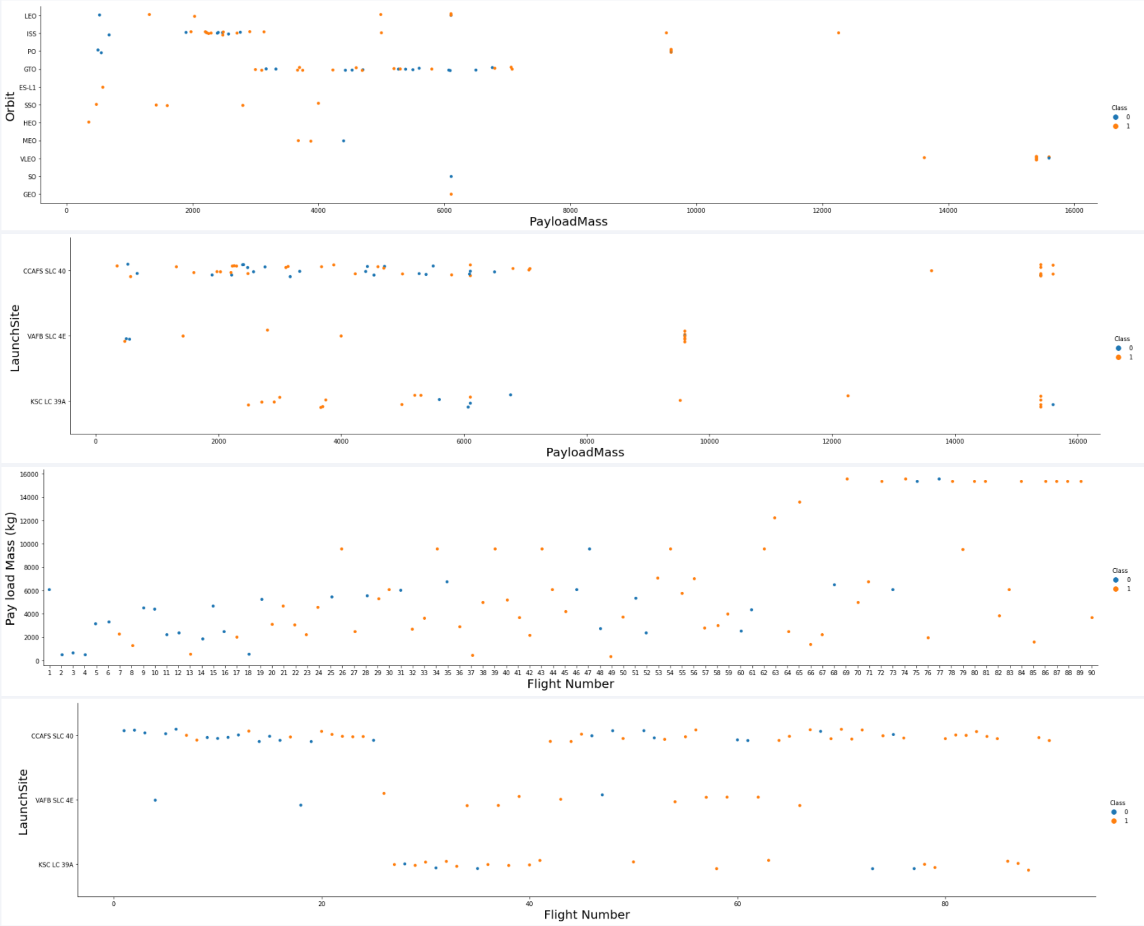
Info to better understand the SpaceX data set (using SQL queries):

- Names of the unique launch sites in the space mission
- Records where launch sites begin with the string 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved.
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass.
- Failed landing outcomes in drone ship, their booster versions, and launch site names in 2015
- Number of landing outcomes between the date 2010-06-04 and 2017-03-20

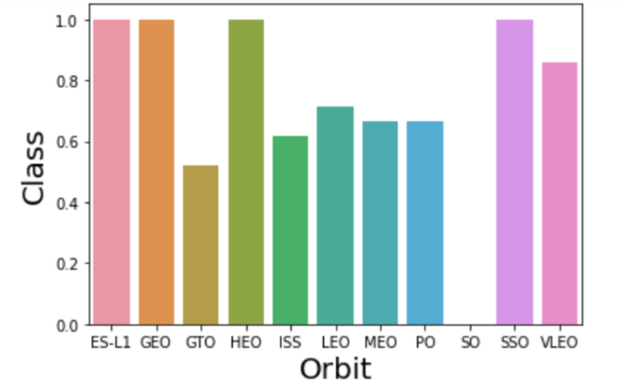
Reference: https://github.com/SchneiderG/Python-for-Data-Science-Labs/blob/main/Capstone/4_SQL_EDA.ipynb

EDA with Data Visualization

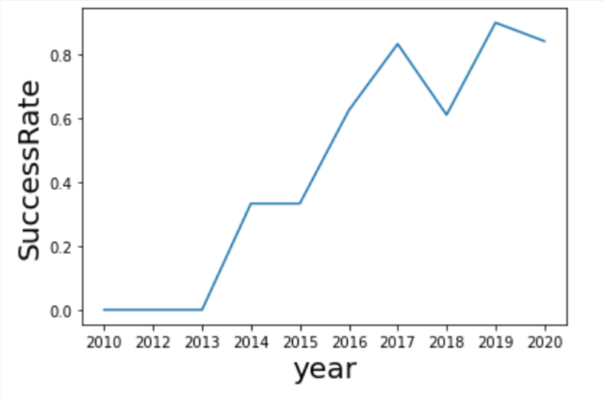
Scatterplots excel at showing relationships between two variables



Bar charts excel at comparing different groups



Line charts excel at showing change over time



Reference: https://github.com/SchneiderG/Python-for-Data-Science-Labs/blob/main/Capstone/5_Python_Data_Visualization.ipynb

Build an Interactive Map with Folium

- Folium builds an interactive map to perform interactive visual analytics. We marked the launch site locations, successful/failed launches for each site on the map and calculated the distances between a launch site to its proximities.
- We can explore the map with those markers and try to discover any patterns from them to choose an optimal launch site.

Reference: https://github.com/SchneiderG/Python-for-Data-Science-Labs/blob/main/Capstone/6_Folium_Maps.ipynb

Build a Dashboard with Plotly Dash

- Building a dashboard application with the Python Plotly Dash package, allows us to find more insights from the SpaceX dataset more easily than with static graphs.
- This dashboard application contains input components such as a dropdown list and a range slider to interact with:
  - Pie chart showing launch successes from certain launch sites
  - Scatter plot showing correlation between payload and success for certain launch sites

Reference: https://github.com/SchneiderG/Python-for-Data-Science-Labs/blob/main/Capstone/7_Plotly_Spacex_Dashboard_app.py

Predictive Analysis (Classification)

Build Model
- Load dataframe
- Standardize data

Evaluate Model
- Split data into Test/Train
- Fit into GridSearchCV

Find Best Classification Method
- Confusion Matrix(s)
- Calculate Accuracy

Reference: https://github.com/SchneiderG/Python-for-Data-Science-Labs/blob/main/Capstone/8_Machine_Learning_Predictions_Accuracy.ipynb

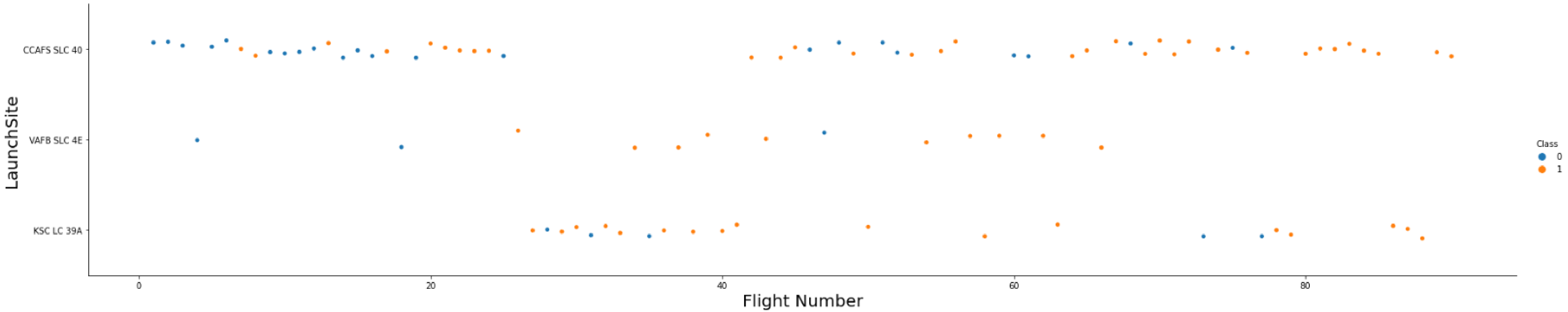
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

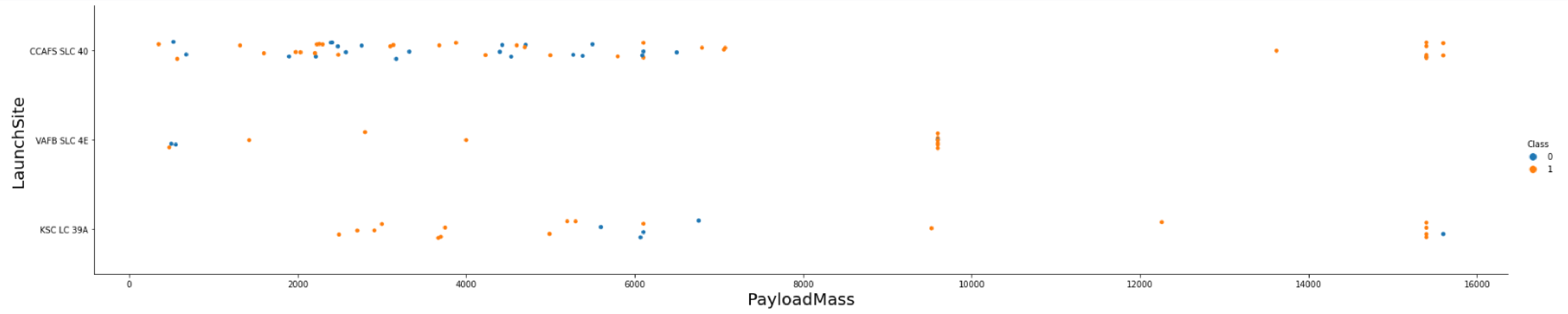
Insights drawn from EDA

Flight Number vs. Launch Site



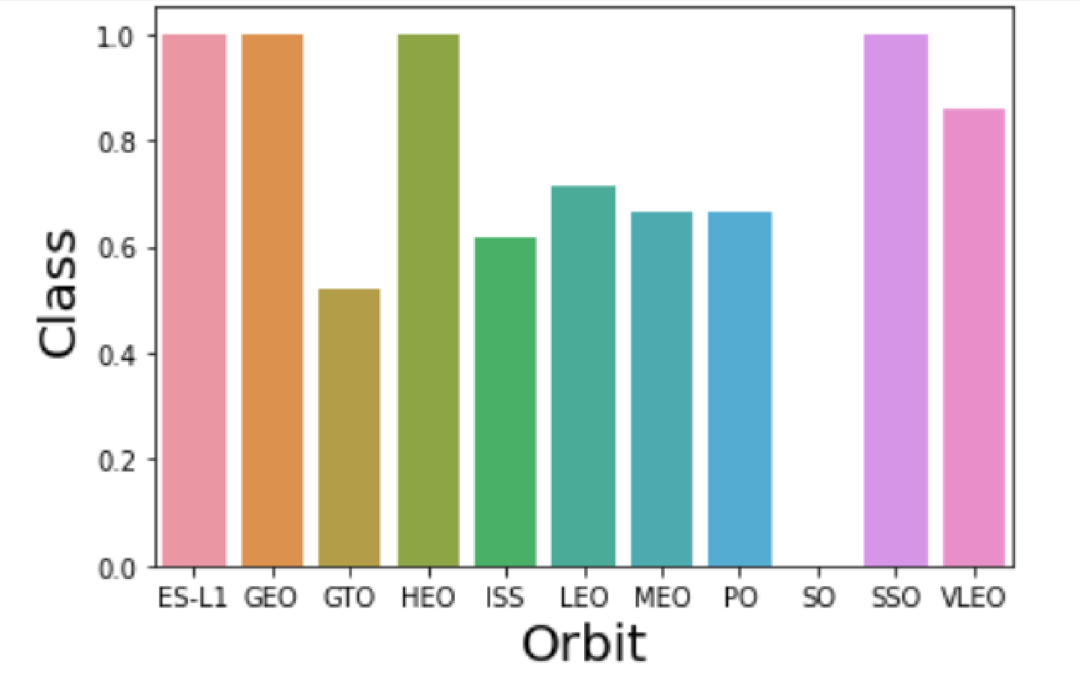
The more flights from a launch site, the more likely it will be successful.

Payload vs. Launch Site



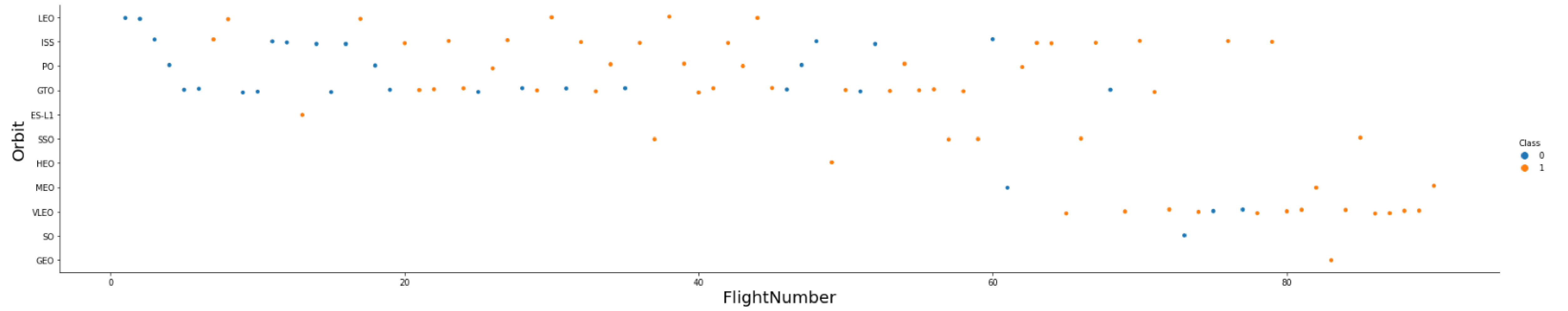
CCAFS launch site has excellent launch success rate with large payloads (>15,000kg).
VAFB launch site has excellent launch success rate with medium payloads(<10,000kg).
KSC launch site has the excellent launch success rate with small (<6,000kg) or large (>15,000kg) payloads

Success Rate vs. Orbit Type



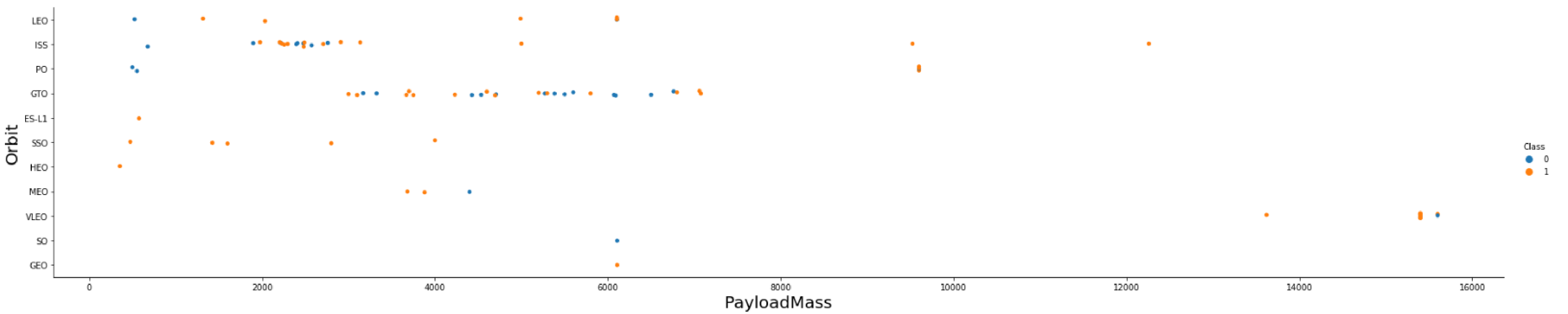
ES-L1, GEO, HEO, SSO have best success rates.

Flight Number vs. Orbit Type



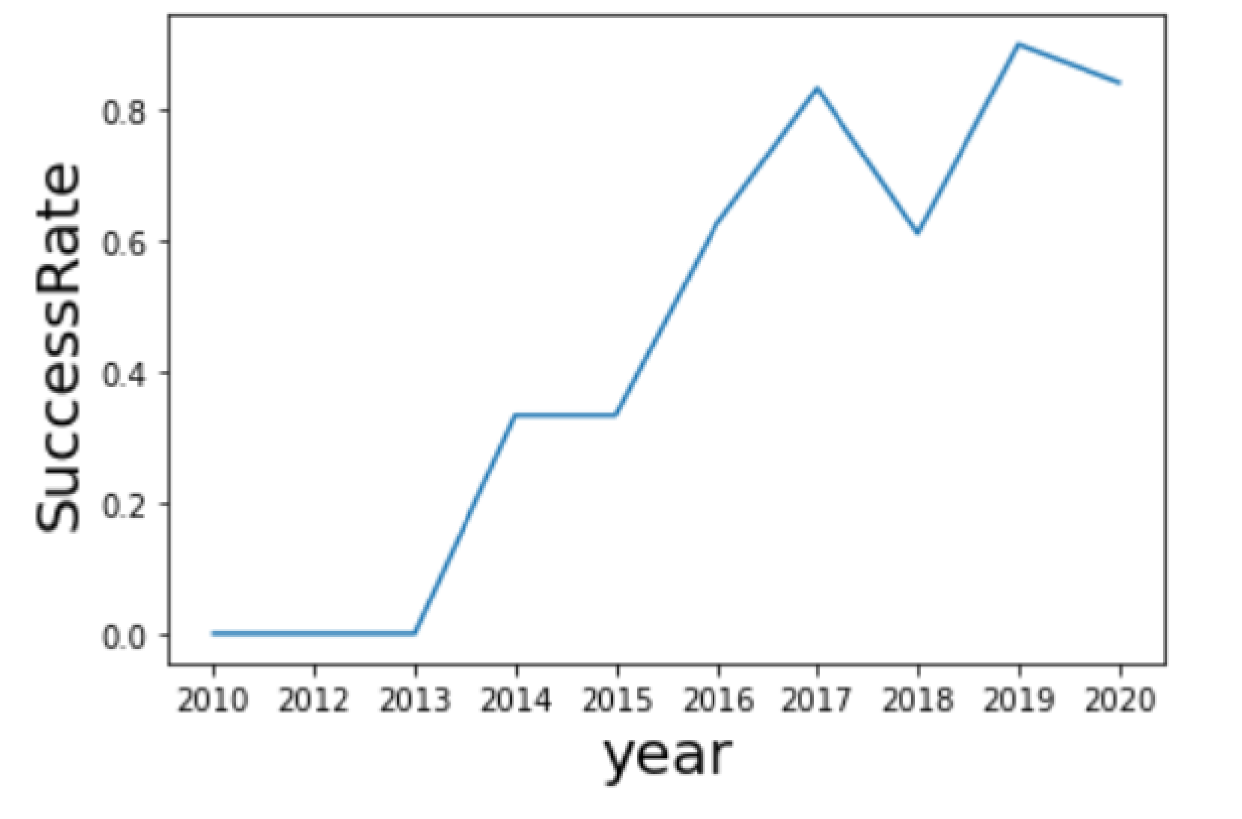
For all orbits, a greater number of flights is positive relationship with a greater number of successful landings.

Payload vs. Orbit Type



Payload can be observed to have a negative relationship with SSO and MEO orbit landing success rate. And a positive relationship with PO, VLEO, and ISS orbit landing success rate.

Launch Success Yearly Trend



Success rate has trended positively after 2013 (although there is dip in 2018).

All Launch Site Names

SQL Query

```
%sql SELECT distinct(launch_site) from SPACEXDATASET
```

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Explanation

Queries the unique launch sites using “DISTINCT”

Launch Site Names Begin with 'CCA'

SQL Query

```
%sql SELECT * from SPACEXDATASET WHERE launch_site like 'CCA%' LIMIT 5
```

Explanation

Queries launch sites beginning with ‘CCA’ using “LIKE” and “LIMITS” it to only 5 entries

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

SQL Query

```
%sql SELECT sum(PAYLOAD_MASS__KG_) from SPACEXDATASET WHERE customer LIKE 'NASA (CRS)'
```

Explanation

Queries the "SUM" of payload mass for only "WHERE" the customer is "LIKE" NASA

totalpayloadmass
45596

Average Payload Mass by F9 v1.1

SQL Query

```
%sql select avg(PAYLOAD_MASS__KG_) as avgpayloadmass from SPACEXDATASET WHERE BOOSTER_VERSION like 'F9 v1.1'
```

Explanation

Queries the “AVG” of payload mass but only for booster versions “LIKE” F9 v1.1

avgpayloadmass
2928

First Successful Ground Landing Date

SQL Query

```
%sql select MIN(date) as "First Successful Ground Landing" from SPACEXDATASET WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

Explanation

Queries the “MIN” date “WHERE” the landing outcome is a successful ground pad landing

First Successful Ground Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query

```
%sql select BOOSTER_VERSION from SPACEXDATASET WHERE LANDING__OUTCOME = 'Success (drone ship)'\
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

Explanation

Queries booster versions "WHERE" landing outcome is a successful drone ship landing and payload mass is between 4000kg and 6000kg

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

SQL Query

```
%sql select MISSION_OUTCOME, count(MISSION_OUTCOME) as "Total" from SPACEXDATASET group by mission_outcome
```

Explanation

Queries the “COUNT” for each mission outcome and “GROUP BY” mission outcome

mission_outcome	Total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

SQL Query

```
%sql SELECT distinct(BOOSTER_VERSION) as "Carried Max" from SPACEXDATASET WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)from SPACEXDATASET)
```

Carried Max
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

Explanation

Queries “DISTINCT” unique booster versions, but only ones that have carried the “MAX” of payload mass

2015 Launch Records

SQL Query

```
%sql SELECT date, booster_version, launch_site from SPACEXDATASET WHERE year(date) = '2015' and landing__outcome = 'Failure (drone ship)';
```

Explanation

Queries date, booster versions, and launch sites in 2015, where the landing outcomes were drone ship failures

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

SQL Query

```
%sql SELECT landing__outcome, count(landing__outcome)as "total" from SPACEXDATASET \
WHERE date between '2010-06-04' AND '2017-03-20' group by landing__outcome order by count(landing__outcome) DESC
```

Explanation

Queries landing outcome "COUNTS"

between 2010-06-04 and 2017-03-20

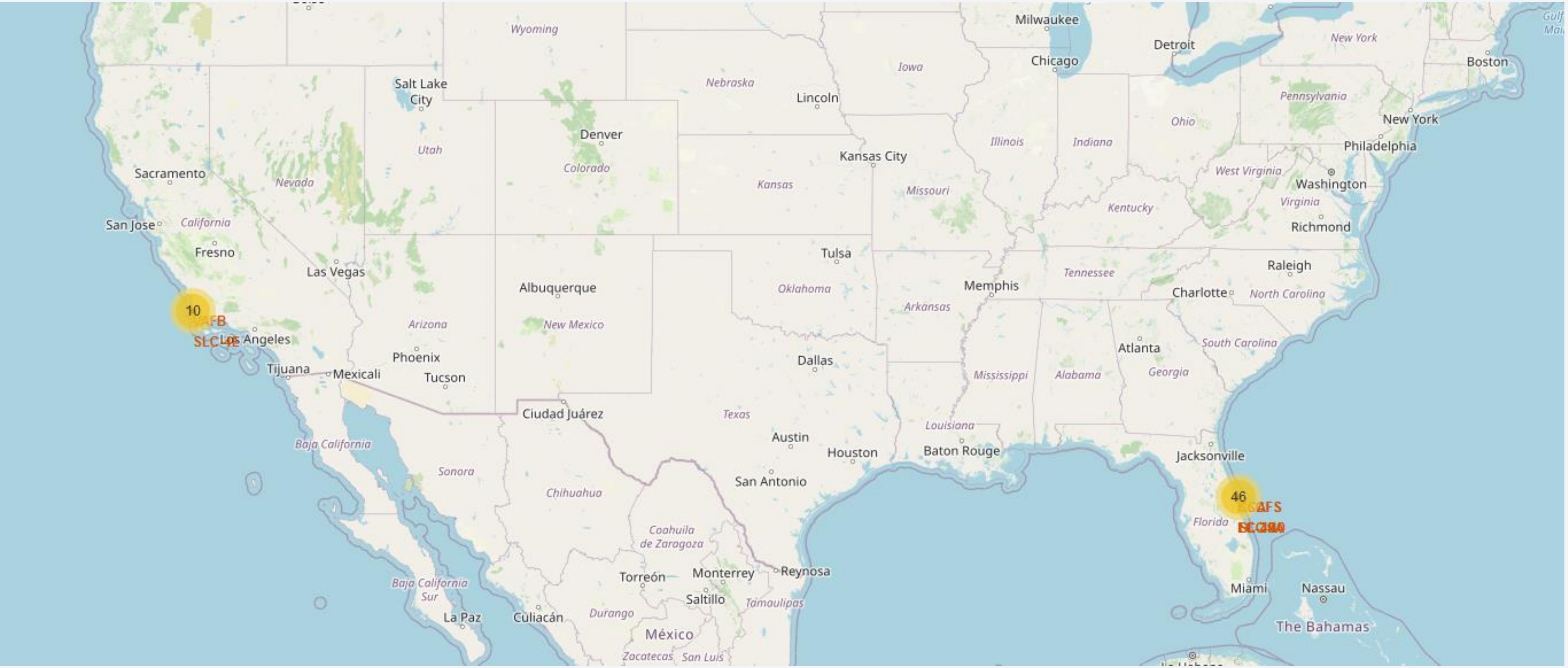
and sorts them in "DESC" order

landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

Launch Site Markers

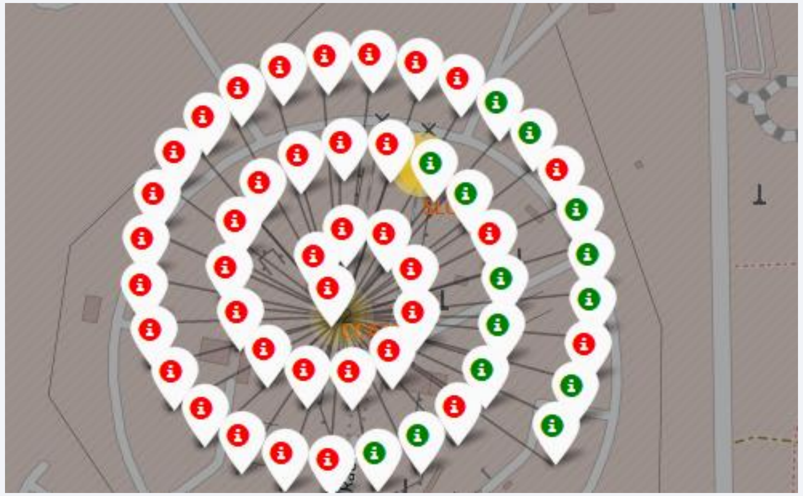


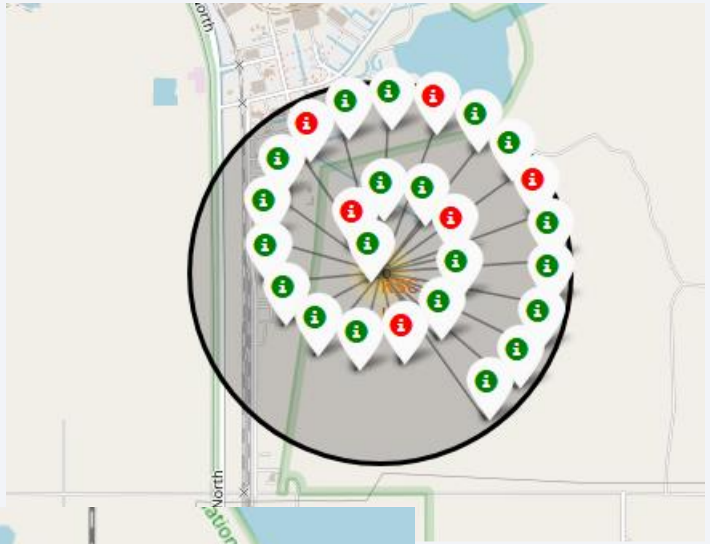
Launch site are within the USA, close to the equator and to the coast

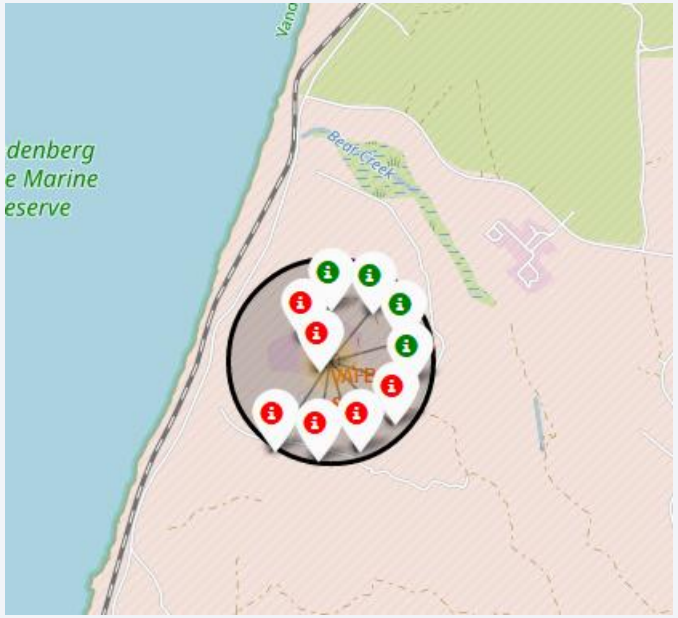
Colored Label Markers

- Colored markets (green for successful launches and red for unsuccessful launches)

- Lists the number of launches per site





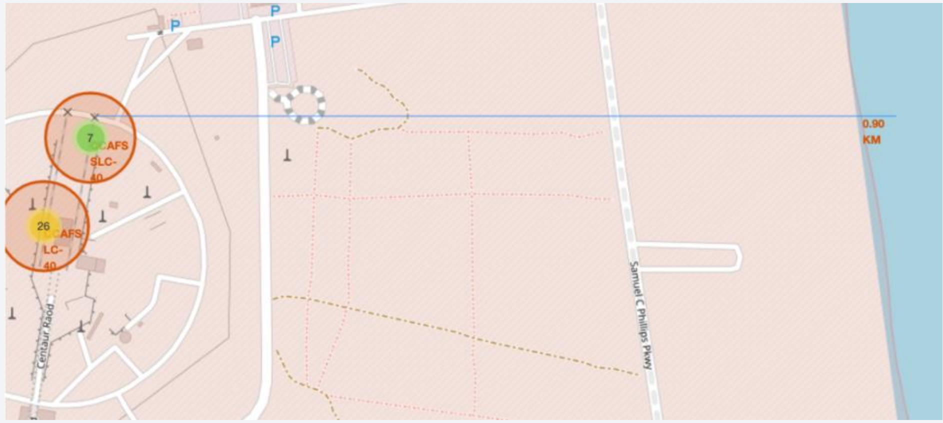


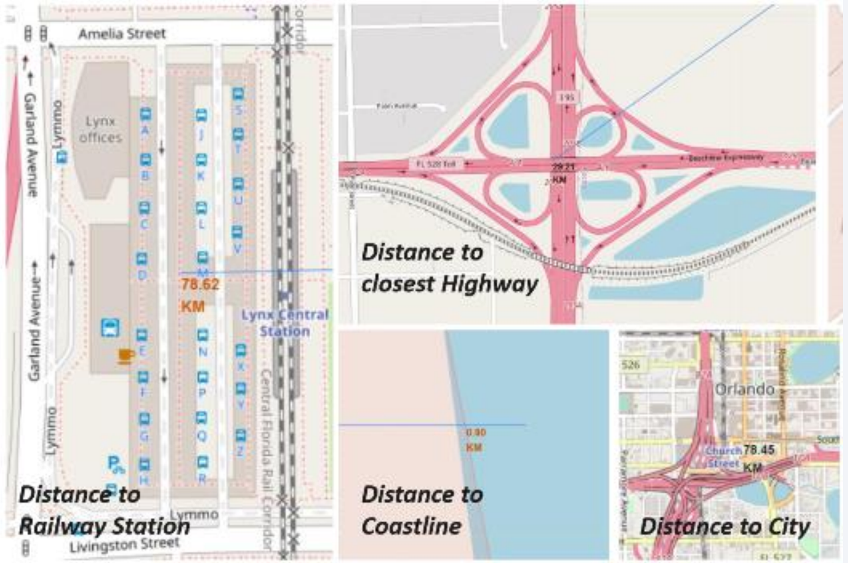
Distance from Launch Sites to Proximities

We draw a line between a launch site to its closest city, railway, highway, etc using MousePosition to find their coordinates on the map first and then calculating distance

Findings

- Are launch sites in close proximity to railways? NO
- Are launch sites in close proximity to highways? NO
- Are launch sites in close proximity to coastline? YES
- Do launch sites keep certain distance away from cities? YES

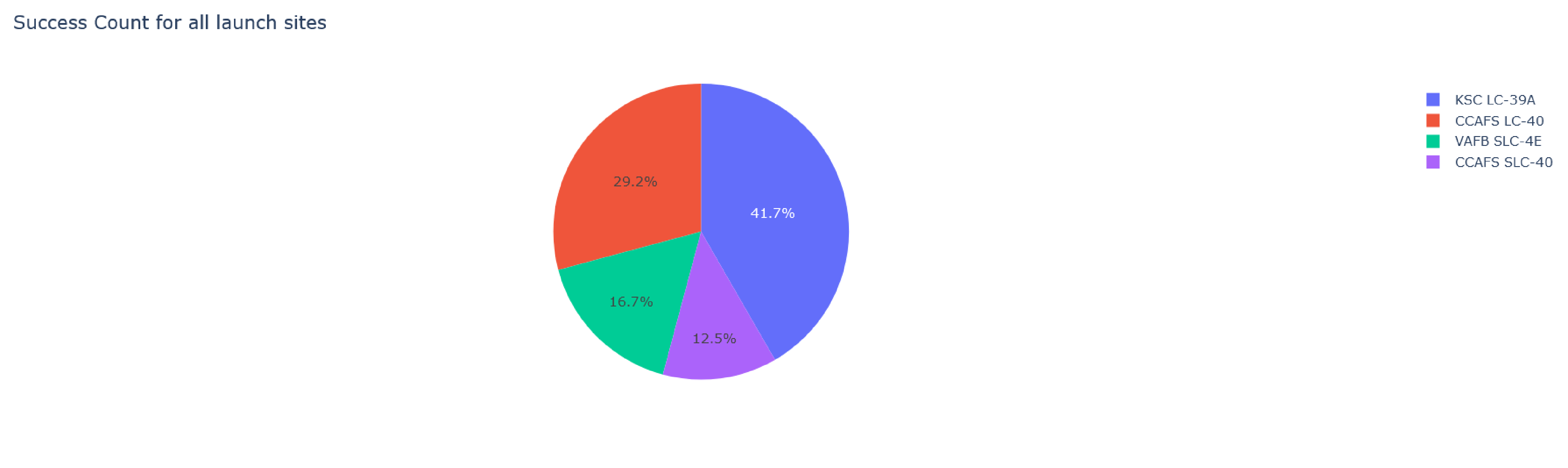




Section 5

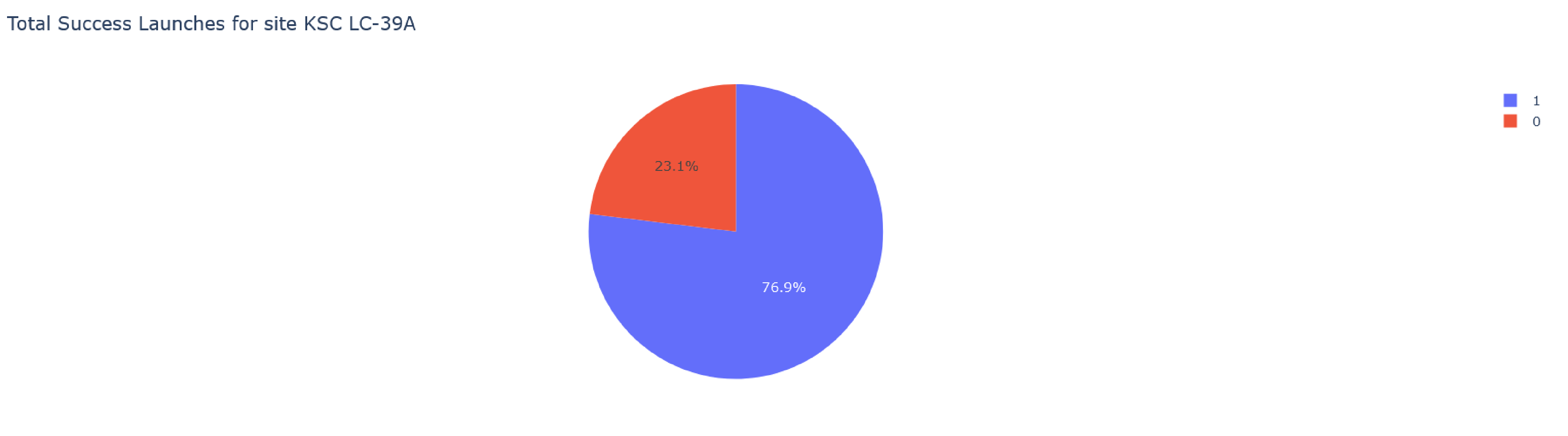
Build a Dashboard with Plotly Dash

Success Count for all Launch sites



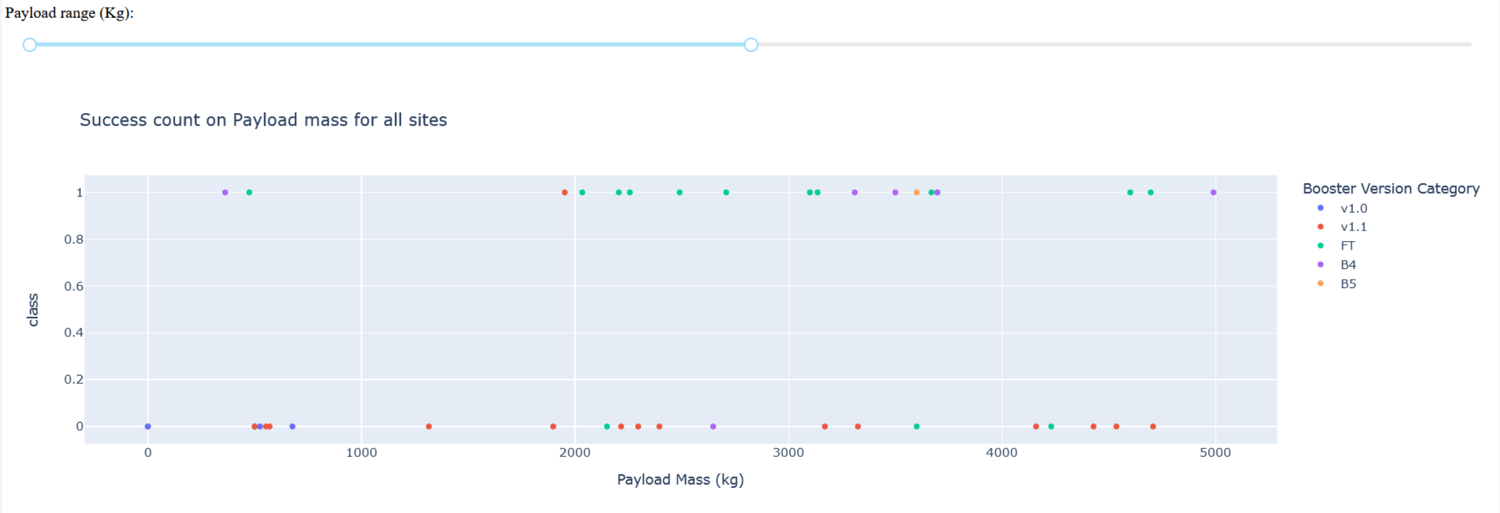
KSC LC-39A Launch site had the most successful launches of all sites

Successful Launches for KSC LC-39A



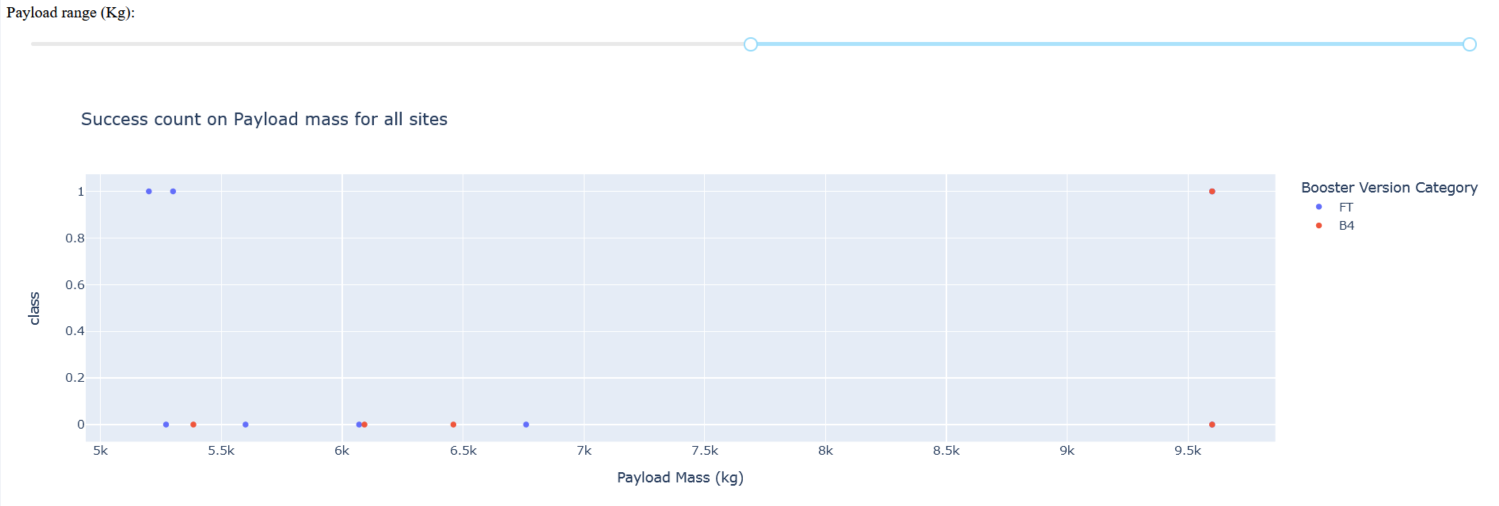
76.9% of launches were a success at KSC LC-39A launch site

Success count on Payload Mass for all sites



- Low Payload Mass (0kg-5000kg)
- FT booster version had the most success
- Overall, lower payload mass launches have more success than high payload mass launches

- High Payload Mass (5,000kg – 10,000kg)
- Very few successes here
- FT and B4 booster types are the only to succeed

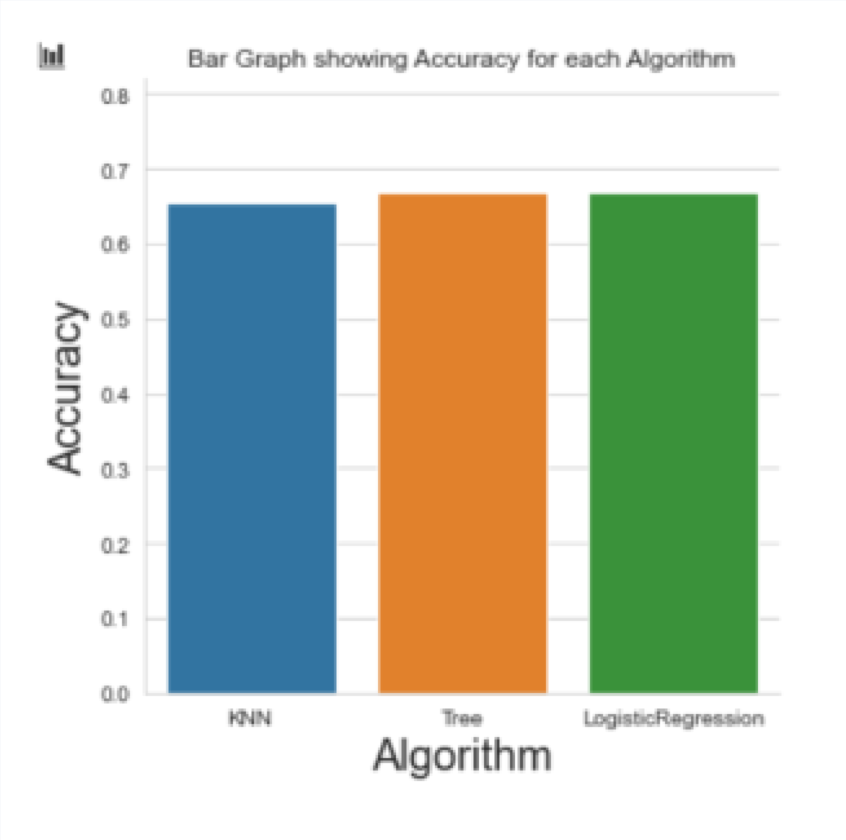


Section 6

Predictive Analysis (Classification)

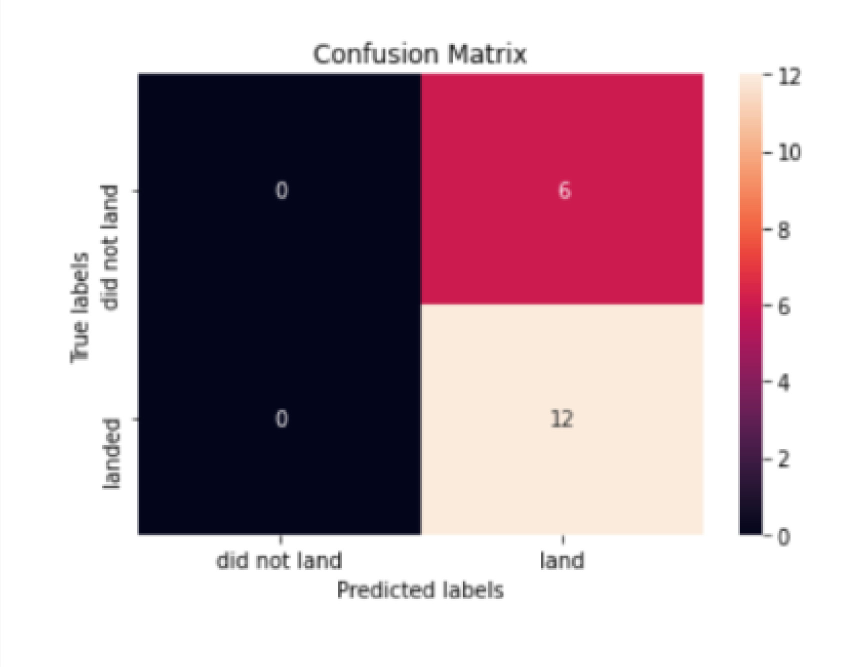
Classification Accuracy

- Visualizing the built model accuracy for all built classification models, in a bar chart shows the decision tree barely wins
- After selecting parameters for the decision tree, we achieve 83.33% accuracy on test data



Confusion Matrix

- Examining the confusion matrix, we see that Decision Tree can distinguish between the different classes. We see that the major problem is false positives.
- Decision Tree is the best performing model



Conclusions – SpaceX Launches

- Success rate of launch's has continued to rise since 2013
- Launch site KSC LC-39A had the most successful launches from all sites
- Low payload mass launches (<5,000kg) have higher landing success rates
- Orbits GEO, HEO, SSO, and ES-L1 had the best success launch rate
- The Decision Tree Classifier Algorithm is best suited for this dataset

Thank you!